Students have to solve the programming tasks and get at least 33% of all points in order to pass.

Team

First you have to form a team. Each team should consist of 3–4 students that attend the *Information Retrieval* course this winter term. Choose a name for your team. Organize the infrastructure that allows all the team members working on the project/source code.

Programming Task

Program your own (command-line based) Information Retrieval system using *Apache Lucene*[1] (at least version *3.6*, currently the newest version is *7.5.0*) and *Java*. Lucene is an open source search library that provides standard functionality for analyzing, indexing, and searching text-based documents. The following criteria have to be met by your Information Retrieval system:

- Using Lucene, parse and index *Plain Text* and *HTML* documents that a given folder and its subfolders contain. List all parsed files.
- Consider the English language and use a stemmer for it (e.g. Porter Stemmer).
- Select an available search index or create a new one (if not available in the chosen directory).
- Make possible for the user to choose the ranking model, Vector Space Model (*VS*) or Okapi BM25 (*OK*).
- Print a ranked list of relevant articles given a search query. The output should contain the 10 most relevant documents with their rank, relevance score, file name resp. title (if it is a HTML-file containing title information) and the path.
- Search multiple fields concurrently (*multifield search*): not only search the document's text (body tag), but also its title.

The program should be written in a way that it is runnable without any corrections or modifications of the java runtime environment or source code! Create a jar-File named IR_P.jar. It should process the input:

```
java -jar IR_P.jar [path_to_document_folder] [path_to_index_folder] [VS/OK] [query]
```

Please send your solutions (jar-File AND source code AND documentation) via e-mail until 20. Dec. 2018, 08:00 a.m. to michael.kotzyba@ovgu.de. In the email, please list all the team members with the team name.

(9 points)

[1] http://lucene.apache.org/java/

Plagiarism of any kind is prohibited! It is not allowed to use any third-party library or source code from others that substantially reduces the complexity of the programming assignment, except an HTML parser (e.g. *Jsoup*) and *Lucene*.

Evaluation for the programming task (9 Points)

Deadline: 20. Dec. 2018, 08:00 a.m.

This list gives a short overview about the scoring of programming assignments.

English		Deutsch	
correctness/completeness	(6)	Korrektheit/Vollständigkeit	(6)
documentation solution (1-2 pages)*	(1)	Doku. Lösungsweg (1-2 Seiten)	(1)
source code quality (structure** and comments)	(2)	Quellcodequalität (Struktur und Kommentare)	(2)

* How did you solved the task (what library function have you used and for what purpose), how does the program work and how can I use it (like user guide)?
Furthermore, list the used version numbers for Lucene and Java, etc.!

** Is the source code comprehensible and structured well (variable names, methods, etc.)?

Note: Plagiarism of any kind is prohibited!